Comparing Differential Item Functioning (DIF) Methods for Assessment

R. Brock Mutcheson[12]; Caitlin Bassett[1]; Tracey Criss[1]; Richard C. Vari[1]
[1]Virginia Tech Carilion School of Medicine; [2]IAMSE Fellowship Participant and ESME Graduate

Background

- Virginia Tech Carilion School of Medicine (VTCSOM) leadership established the **InclusiveVTCSOM Task Force** to highlight important diversity-related opportunities within admissions, the curriculum, and the learning environment.
- The purpose of this analysis was to strengthen assessment item quality review rigor by proactively evaluating trends in performance data for evidence of racial, ethnic, or gender bias. Four methods for detecting bias were compared.

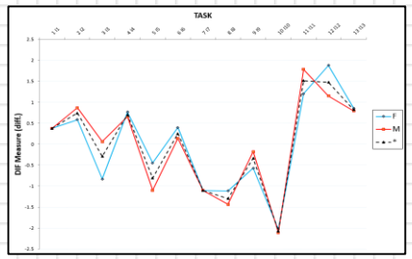
Methods

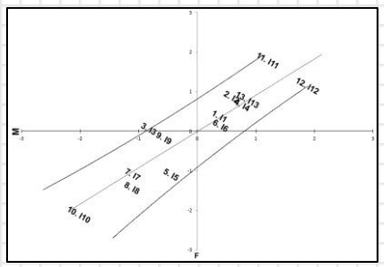
- We constructed instrument validity arguments using a unified approach (Cook, 2015).
- We aggregated five (5) years of M3 summative assessments, including ratings and comments (n=268).
- We linked nine (9) different courses with self-reported individual demographic data.
- We coded each demographic variable of interest (i.e. gender; racial or ethnic minority) dichotomously based on self-identification.
- We assessed Uniform and Non-uniform differential item functioning (DIF) in Winsteps and Stata 16 using Rasch, logistic regression, and Mantel-Haenszel.
- Effect sizes were interpreted using Zwick (1999) standards for MH and Zumbo (1999) standards for LR to determine whether items should be flagged for review. Level A = negligible; Level B = slight to moderate; Level C = moderate to large.
- We tabulated performance-based adjectives for each group in Nvivo. Each adjective fit one category: *standout, grindstone, ability, or compassion (*Ross et al., 2017).

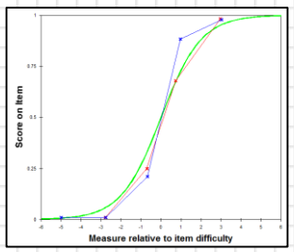
Results

Potential Differential Item Functioning

Method	Type	Self-Identification: Female		Self-Identification: Minority	
		(%)	Repeat	(%)	Repeat
1: Mantel-Haenszel		1%	0%	2%	0%
2: Logistic Regression	Uniform	5%	0%	4%	2%
	Non-Uniform	5%	1%	5%	3%
3: Rasch	Uniform	0%	0%	0%	0%
	Non-Uniform	2%	1%	1%	0%







Item	Nonuniform		Uniform	
	Chi2	Prob.	Chi2	Prob.
q1	0.31	0.57	0.10	0.76
q2	0.05	0.83	0.14	0.71
q3	0.01	0.92	4.71	0.03
q4	1.90	0.17	1.21	0.27
q5	4.28	0.04	0.25	0.61
q6	0.21	0.65	0.28	0.60
q7	1.06	0.30	0.00	0.98
q8	0.53	0.47	0.38	0.54
q9	2.16	0.14	1.14	0.28
q10	4.86	0.03	0.24	0.63
q11	0.10	0.75	2.54	0.11
q12	1.44	0.23	2.39	0.12
q13	0.05	0.83	0.91	0.34

Virginia Tech-IRB-21-339

Conclusions

- Differential item functioning (DIF) methods are the first step in detecting the extent of potential item bias.
- Advantages and disadvantages to each DIF method should be understood and the corresponding assumptions should be tested prior to analysis.
- Logistic regression analysis detected more items with potential DIF than other methods. However, practical significance was low in all cases. Moreover, expert item review suggested these cases arose by chance or were caused by confounding.
- Qualitative comparisons of the adjective weighted frequencies provided additional nuance in understanding evaluation use.

Discussion

- Items with statistically significant DIF should be subsequently evaluated using expert item review. Moreover, DIF magnitude and the clinical or practical significance should be considered (Scott, et. al., 2010) before final decisions are made about an item.
- Graphical representations and qualitative approaches can aid in interpreting DIF. For example, narrative evaluations can contain bias by focusing on stereotypes of certain groups based on race, ethnicity, gender, or other characteristics.